Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 24 (2016) 1074 – 1079

Procedia Technology

International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)

Single Channel Speech Separation With Frame-based Summary Autocorrelation Function Analysis

Raghi E.R[a],Lekshmi M.S[b]

[a]*PG Scholar,Ilahia College of Engineering and Technology,Ernakulam,686673,India*
[b]*Asst.Professor.Ilahia College of Engineering and Technology,Ernakulam,686673,India*

Abstract

Single channel speech separation system with frame based pitch range estimation is presented. The system decomposes the input mixture speech into frames and pitch of the speech is estimated in each frame of modulation spectrum by using summary autocorrelation function (SACF) analysis. In this method, each frame is divided into two, above and below 1 KHz and compute generalized autocorrelation for periodicity detection. By using this pitch range, a mask is created to filter the target speech from noisy mixture speech. Performance evaluation of the proposed system shows a better response compared to the existing methods.

Peer-review under responsibility of the organizing committee of ICETEST – 2015

*Keywords:*Summary autocorrelation function; Modulation spectrum; Periodicity detection.

1. Introduction

In all listening situations, sound reaching our ear is composed of target speech and other background noise from multiple sources. Performance of the automatic speech recognition, telecommunication and audio retrieval systems depend on the quality of speech entered into that system. In order to improve the quality of the speech, the speech separation system is often required in the pre-processing stage of many applications. Many works have been done in the single channel speech area for preserving quality of the target speech [1, 2, 3].From these works, computational auditory scene analysis (CASA)methods made a special attraction to separate target speech from acoustic mixture. CASA methods are based on Auditory Scene Analysis (ASA) principle. The auditory system of human partitioned the incoming auditory information into streams corresponding to different sources. This process of segregating

* Corresponding author. Tel.: 9400407280
.E-mail address:raghi.er@gmail.com

Peer-review under responsibility of the organizing committee of ICETEST – 2015
doi:10.1016/j.protcy.2016.05.239

auditory scene into separate mental representations is known as auditory scene analysis (ASA).Inspired from ASA principles, many researchers were devoted in constructing computational auditory scene analysis (CASA) system for speech separation[4,5].Wang and Brown model is such an example [6] which is based on the oscillatory correlation. But the system is not capable of handling the unresolved portion of the speech.Hu and Wang model [7] employed different method to segregate resolved and unresolved harmonics of the target speech and but the model is limited to segregate only the voice speech. Mahmoodzadeh model [8] uses onset and offset for pitch estimation. This system decomposes the acoustic mixture into time-frequency frames and create frequency mask for removing the interference part of the mixture. All these techniques require accurate pitch estimation method. Accuracy of the speech separation system is mainly concentrated in pitch estimation method. This paper proposes frame based pitch range estimation that estimates the pitch range of both target and interference portion of the speech by using summary autocorrelation function (SACF).And by using this pitch range, system mask out the interference portion from the mixture. This paper is organized as follows. In section 2, we first give a brief description of our system and then present the details of each stage. The results of the system are reported in section 3. The paper concludes with a discussion in section 4.

2. System Description

The main idea of the system is to remove the interference portion of the mixture signal to extract the target speech. Thereupon, at first modulation frequency of the acoustic mixture signal is computed and the pitch range of both target and interference speech is estimated with the help of SACF. Finally, by using this pitch range, a mask is created to separate the target speech. The block diagram of the proposed system is shown in fig.1. The system contains mainly four steps: T- F decomposition, modulation transform, pitch range estimation and speech separation. And the detailed description of the each block is as follows.

2.1 T-F Decomposition

The acoustic noisy input is a broadband signal. For the analysis, we want to convert this broadband signal into narrowband subband signal. At the T-F decomposition stage, input mixture signal is decomposed into narrowband signal. Here Short Time Fourier Transform (STFT) is used for decomposition. In STFT, short time signal is obtained by windowing the long time speech signal. Then Fast Fourier Transform (FFT) of each windowed segment is computed.

$$X(m, k) = STFT\{x[n]\} \quad (1)$$

$$= \sum_{n=0}^{K-1} (x(n) w(mM - n) \exp(-j2\pi nk/K)$$

Where K is the DSTFT length, w (·) is the acoustic frequency analysis window with length L and M is the decimated factor. So after STFT, wideband input speech signal is divided into number of channels or frames.

2.2 Modulation Transform

After the T-F decomposition, each channel can be represented as the multiplication of two processes: a high frequency carrier and low frequency modulator. Narrowband frequency subband from the T-F decomposition module is divided into carrier and modulator signal. That is,

$$X(m, k) = M(m, k)\, C(m, k) \quad (2)$$

The modulator signal M(m,k) is obtained by applying an envelope detector to each channels, as

$$M(m, k) \triangleq D\{X(m, k)\} \quad (3)$$

Where D is the operator of the envelope detector. This modulator signal is used for further processing. Then, the time index m in the modulator signal is transformed into frequency index by taking Fourier transform.

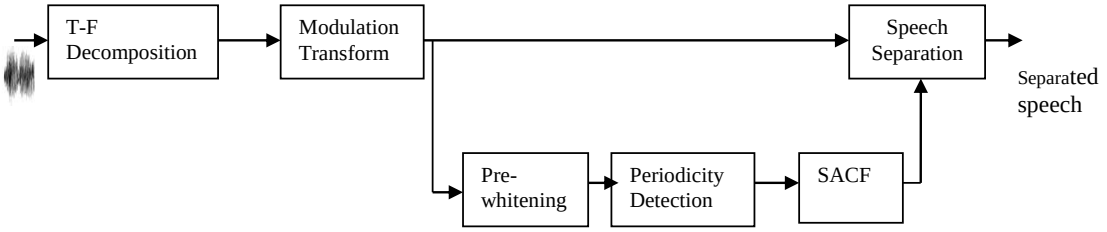

Fig. 1.Basic block diagram of the proposed system

2.3 Pitch Range Estimation

Here pitch range of target and interference speech is obtained by computing the steps; pre-whitening, periodicity detection, and calculation of SACF.

2.3.1 Pre-Whitening

Most of the computational auditory scene analysis systems further process each frames in order to extract feature that are useful for grouping. Here each modulation frequency frame of mixture signal from the previous stage is processed by pre-whitening filter. Pre –whitening removes the short – time correlation of the noisy signal. The pre-whitening filter is implemented by using warped linear prediction (WLP).

2.3.2 Periodicity Detection

The correlogram is computed in modulation frequency domain by performing an autocorrelation at the output of each channel. correlogram detects the periodicities present in the each frames. For this, firstly pre-whitened modulation frequency frames of mixture signal are splitting in to two bands, below and above 1 kHz. This is done with low pass and high pass filters. Thereby, resolved and unresolved part of the signal is separated and is treated differently. These filters have 12 dB/octave attenuation in stop band .Next step is half-wave rectification of high frequency signal. To obtain envelope, resulting high frequency signal is lowpass filtered. A generalized autocorrelation is then computed for the low-frequency band and the envelope of the high frequency band for periodicity detection. The autocorrelation can be performed in the modulation frequency domain by means of the Discrete Fourier Transform (DFT) and its inverse transform (IDFT). Speed of the computation is increased by using Fast Fourier Transform (FFT) and its inverse (IFFT).

2.3.3 SACF

The summation of the generalized autocorrelation of both high and low frequency frames corresponds to the SACF. That is,

$$x_2 = IDFT(|DFT(x_{low}|)^k + IDFT(|DFT(x_{high}|)^k$$

$$= IDFT(|DFT(x_{low}|^k + |DFT(x_{high}|^k) \quad (4)$$

Where k determines the frequency domain compression and its value is smaller than 2. The SACF exhibits peaks at the period of each fundamental frequency. So the highest peak in resulting SACF signal corresponds to the pitch ofthe dominant speech. So the highest and lowest onset position detection determines the pitch range of both target and interference speech.

2.4 Speech Separation

After finding pitch range of target and interference speaker, system separates the target speech by using frequency masking. Here the system masked out the interfering speaker. For generating frequency mask, first we have to evaluate the mean of modulation spectral energy over the pitch frequency of both the target and interference signals. They can be represented as,

$$E_t^k = \frac{\sum_{i \epsilon Q^k} (|X(k,i)|)^2}{target\ pitchrange} \quad (5)$$

$$E_I^k = \frac{\sum_{i \epsilon Q^k} (|X(k,i)|)^2}{Interferance\ pitchrange} \quad (6)$$

Then compare the modulation spectral energy of the target and interference speakers by using this equation,

$$F^k = \frac{E_t^k}{E_t^k + E_I^k} \quad (7)$$

The resulting the frequency masking function is not applied in the modulation frequency domain directly. There are artifacts associated with it. So, frequency mask is transformed into the time domain by taking inverse FFT.ie,

$$f^k(m) = IFFT(F^k) \quad (8)$$

Then the speech is separated by convoluting the obtained filter $f^k(m)$ with the modulator signal of the mixture signal. And then, original target speech is reconstructed by multiplying the carrier signal.ie,

$$\tilde{X}(m,k) = [M(m,k) * f^k(m)]C(m,k) \quad (9)$$

From this target signal in the modulation frequency domain is obtained. To get separated target signal in the time domain, take the inverse STFT of $\tilde{X}(m,k)$.

3. Experimental Results

A series of experiments have been conducted to evaluate the accuracy of the proposed method. We have taken samples speech from TIMIT database and mixed them linearly to get the mixture. The mixtures are generated in such a way that one of the speeches is dominant. Matlab program of both Mahmoodzadeh model and proposed model are simulated with different type of mixture and the proposed system was compared with the Mahmoodzadeh model. The database used in the Mahmoodzadeh model in [8] is similar to the database considered in the proposed system. Result obtained by simulated Mahmoodzadeh model is comparable with that of the Mahmoodzadeh model in [8]. Separation performance was measured with signal-to-noise ratio (SNR), Perceptual Evaluation of Speech Quality (PESQ), and the overall quality (OVRL). The SNR is defined as:

$$SNR = 10\ log_{10} \frac{\sum_n x^2(n)}{\sum_n [x(n) - \check{x}(n)]^2} \quad (10)$$

Where *x(n)* is the original speech and *$\check{x}$(n)* is the separated speech. Table 1 shows the SNR of mixture speech and separated speech of Mahmoodzadeh model and proposed model. The measure, overall quality (OVRL) is obtained by linearly combining the Perceptual Evaluation of Speech Quality (PESQ), Log-likelihood ratio (LLR), and Weighted Spectral Slope (WSS) measures. Where, LLR is Log-likelihood ratio obtained from LPC analysis of original and separated speech. WSS is a direct spectral distance measure and based on comparison of smoothed spectra from the clean and distorted speech samples. Table 2 and 3 show PESQ and overall quality of Mahmoodzadeh model and proposed model respectively. All results show that the proposed system yields better performance and the SNR of segregated speech is improved from the mixture. From the Welch power spectral density of separated speech in fig.2, it is clear that the dominant speech can be separated without loss of information.

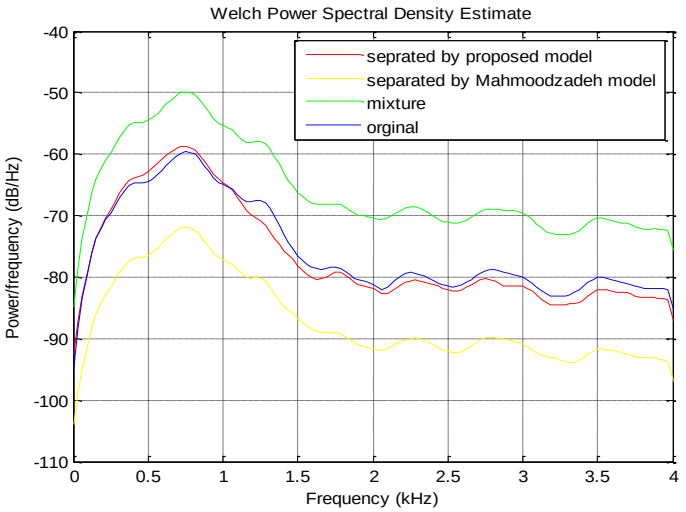

Fig.2.Welch power density of original, mixture and separated speech

A- Mixture of Two female speakers.
B- Mixture of Two male speakers.
C- Mixture of male & female speakers with female dominant.
D- Mixture of male & female speakers with male dominant.

Table 1.SNR results for separated and original mixtures

Type of mixture	Signal to Noise Ratio(dB)		
	Mixture	Mahmoodzadeh model	Proposed system
A	-6.6295	2.255	7.8569
B	-6.4927	3.4619	9.3763
C	-6.1279	2.1866	7.7758
D	-8.5507	1.3483	4.9985

Table 2.PESQ of segregated speech of mahmoodzadeh model and proposed model

Type of mixture	PESQ	
	Mahmoodzadeh model	Proposed system
A	2.3142	2.3736
B	2.5326	2.9021
C	2.5204	2.5339
D	2.6588	2.7578

Table 3.Overall Quality of segregated speech of mahmoodzadeh model and proposed model

Type of mixture	Overall Quality	
	Mahmoodzadeh model	Proposed system
A	1.4446	4.0110
B	4.3196	4.4579
C	4.0355	4.0908
D	3.0810	4.1378

4. Conclusion and Discussion

In this paper, we presented a new single channel speech separation system that estimates the pitch range in each frames by analyzing summary autocorrelation function. It results good estimate of pitch range to produce frequency mask and eliminate the interference portion from the mixture. From the signal to noise ratio, PESQ and overall quality, it is clear that the proposed system is superior to the existing Mahmoodzadeh model.

References

[1] G Hu, D Wang, "A Tandem algorithm for pitch estimation and voicedspeech segregation," *IEEE Trans Audio Speech Lang Process*. 18(8), 2067–2079 (2007).

[2] JJ. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model,"*IEEE Trans. Speech and Audio Process,* vol. 9, pp. 731-740, 2001.

[3] P. Varga and R. K. Moore, "Hidden Markov model decomposition ofspeech and noise," Proceedings of ICASSP, pp. 845-848, 1990. NJ, 2006).

[4] Brown GJ, Wang DL (eds.), Computational Auditory Scene Analysis: Principles, Algorithms, and Applications (Wiley & IEEE, Hoboken, *Process*. 18, 2991–3002 (2005)

[5] M Buchler, S Allegro, S Launer, N Dillier, "Sound classification in hearing aids inspired by auditory scene analysis",*EURASIP J Appl Signal*

[6] D.L. Wang and G.J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation*," IEEE Trans. Neural Net.*, Vol. 10, pp. 684-697, 1999.

[7] G. Hu, and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks,* vol. 15, no. 5, pp. 1135-1150, 2004.

[8]A.Mahmoodzadeh, H.R Abutaledi ,H.Soltanian,H.Sheikhzadeh,"Single channel speech separation with a frame – based pitch range estimation method in modulation frequency,"*IEEE Trans Audio Speech Lang Process*,2012.